PATENT ABSTRACTS OF JAPAN

(11)Publication number : **2003-022088**
(43)Date of publication of application : **24.01.2003**

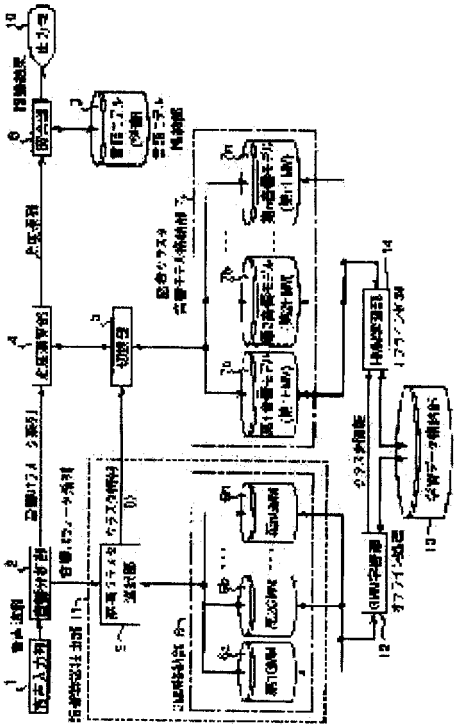
(51)Int.Cl. **G10L 15/06**
G06F 3/16
G10L 15/02
G10L 15/10
G10L 15/14
G10L 15/18

(21)Application number : **2001-209503**
(22)Date of filing : **10.07.2001**

(71)Applicant : **SHARP CORP**
(72)Inventor : **YAMAGUCHI KOICHI**

(54) DEVICE AND METHOD FOR SPEAKER'S FEATURES EXTRACTION, VOICE RECOGNITION DEVICE, AND PROGRAM RECORDING MEDIUM

(57)Abstract:
PROBLEM TO BE SOLVED: To precisely extract speaker's features from a less amount of speech data.
SOLUTION: A GMM learning part 12 adds the value of a vocal tract length expansion/contraction coefficient α to voice data of respective learners stored in a learning data storage part 13, clusters the learners according to the vocal track expansion/contraction coefficient α, performs data conversion so that voice data of a speaker in a cluster nearby some cluster C belong to the cluster C, and reclusters the learners by using GMMs of the respective clusters. The GMMs of the obtained (n) clusters are stored in a GMM storage part 6. A speaker cluster selection part 3 makes the (n) GMMs stored in the GMM storage part 6 operate on a sound parameter series from a sound analysis part 2 and outputs the index of the GMM giving the maximum likelihood as speaker cluster information. Thus, speaker's features are precisely extracted from a small amount of learning data without depending upon speech contents.



LEGAL STATUS

[Date of request for examination] 18.06.2004

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2003－22088
（P2003－22088A）
(43)公開日　平成15年１月24日(2003.1.24)

(51)Int.Cl.[7]	識別記号	ＦＩ		テーマコード（参考）
Ｇ１０Ｌ　15/06		Ｇ０６Ｆ　3/16	３２０Ｇ	５Ｄ０１５
Ｇ０６Ｆ　3/16	３２０		３２０Ｈ	
		Ｇ１０Ｌ　3/00	５２１Ｖ	
Ｇ１０Ｌ　15/02			５１５Ｅ	
15/10			５２１Ｓ	

審査請求　未請求　請求項の数12　ＯＬ　（全 16 頁）　最終頁に続く

(21)出願番号　特願2001－209503(P2001－209503)

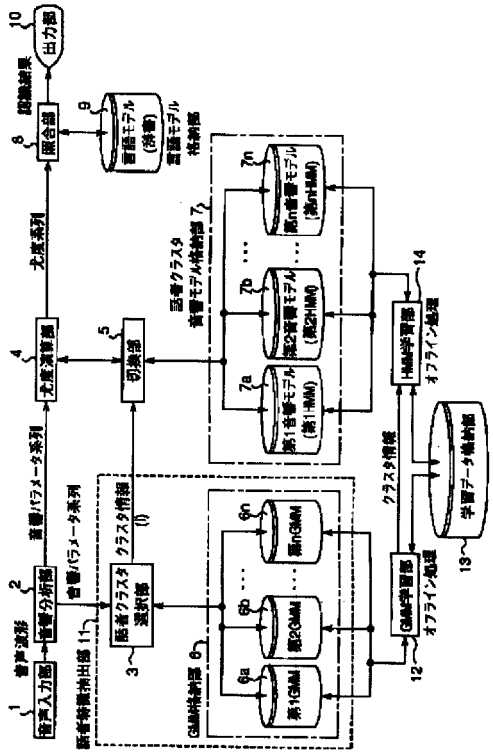
(22)出願日　平成13年７月10日(2001.7.10)

(71)出願人　000005049
シャープ株式会社
大阪府大阪市阿倍野区長池町22番22号
(72)発明者　山口　耕市
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内
(74)代理人　100062144
弁理士　青山　葆　（外１名）

Ｆターム(参考)　5D015 FF04 GG04 HH04

(54)【発明の名称】　話者特徴抽出装置および話者特徴抽出方法、音声認識装置、並びに、プログラム記録媒体

(57)【要約】
【課題】　より少ない発声データから精度良く話者特徴を抽出する。
【解決手段】　ＧＭＭ学習部１２は、学習データ格納部１３に格納された各学習話者の音声データに声道長伸縮係数αの値を与え、この声道長伸縮係数αに基づいて学習話者をクラスタリングし、あるクラスタＣの近傍のクラスタＤに属する話者の音声データを上記クラスタＣに属するようにデータ変換し、各クラスタのＧＭＭを用いて学習話者をクラスタリングし直す。得られたｎ個のクラスタのＧＭＭはＧＭＭ格納部６に格納される。話者クラスタ選択部３は、音響分析部２からの音響パラメータ系列にＧＭＭ格納部６に格納されたｎ個のＧＭＭを作用させ、最大尤度を与えるＧＭＭのインデックスを話者クラスタ情報として出力する。こうして、少ない学習データから、発話内容に因らずに精度良く話者特徴を抽出する。



1

【特許請求の範囲】

【請求項１】　入力話者の音声に基づいて、標準話者の音声と上記入力話者の音声との関係を表わすパラメータを話者特徴として抽出する話者特徴抽出装置において、

各学習話者に関して、上記標準話者に対する声道長の伸縮係数αを所定の方法によって予め求め、この求められた伸縮係数αの値に基づいて上記学習話者をクラスタリングする学習話者クラスタリング手段と、

上記クラスタリングされた各クラスタに属する話者集合毎に、学習によって１状態の混合ガウス分布型音響モデルを生成する音響モデル生成手段と、

上記生成された１状態の混合ガウス分布型音響モデルの夫々に対する上記学習話者の音声サンプルの尤度を算出し、その尤度に基づいて上記学習話者を再クラスタリングする再クラスタリング手段と、

上記音響モデル生成手段と再クラスタリング手段とを制御して、所定の条件を満たすまで、上記１状態の混合ガウス分布型音響モデルの生成と上記学習話者の再クラスタリングとを繰り返すループ学習手段と、

上記ループ学習手段によって最終的に生成された１状態の混合ガウス分布型音響モデルの群を格納する音響モデル格納部と、

上記音響モデル格納部に格納された１状態の混合ガウス分布型音響モデルの夫々に対する入力話者の音声サンプルの尤度を算出し、最大の尤度を呈する１状態の混合ガウス分布型音響モデルを入力話者が属するクラスタの情報として選択する話者クラスタ選択部を備えて、

上記入力話者の特徴として上記最大の尤度を呈する１状態の混合ガウス分布型音響モデルを抽出することを特徴とする話者特徴抽出装置。

【請求項２】　請求項１に記載の話者特徴抽出装置において、

上記学習話者クラスタリング手段によってクラスタリングされた各クラスタのうちの注目クラスタに隣接する隣接クラスタに属する学習話者の音声サンプル、または、上記注目クラスタと上記伸縮係数α値の差が所定値以内の近傍クラスタに属する学習話者の音声サンプルに対して、上記注目クラスタと隣接クラスタまたは近傍クラスタとの上記伸縮係数α値に基づいて周波数伸縮を行うことによって、上記注目クラスタに属する音声サンプルを生成し、この生成された音声サンプルを上記注目クラスタに編入して当該クラスタの音声サンプル数を豊富化する操作を、上記学習話者クラスタリング手段によってクラスタリングされた総てのクラスタについて実行する音声サンプル豊富化手段を備えて、

上記音響モデル生成手段は、上記音声サンプル豊富化手段によって音声サンプル数が豊富化された後の各クラスタ毎に、上記１状態の混合ガウス分布型音響モデルを生成するようになっていることを特徴とする話者特徴抽出装置。

2

【請求項３】　請求項２に記載の話者特徴抽出装置において、

上記注目クラスタに属する音声サンプルを生成する際に、上記音声サンプル豊富化手段が上記学習話者の音声サンプルに対して周波数伸縮を行う音声区間を、有音・無音の別および調音点に基づいて限定するようにしたことを特徴とする話者特徴抽出装置。

【請求項４】　請求項１あるいは請求項２に記載の話者特徴抽出装置において、

上記再クラスタリング手段によって上記学習話者を再クラスタリングする場合に、上記再クラスタリングの対象となる対象学習話者が再クラスタリングの前に属していたクラスタの伸縮係数αと再クラスタリング後に属するクラスタの伸縮係数αとが所定値以上離れている場合には、当該対象学習話者を上記再クラスタリングの対象から外すようになっていることを特徴とする話者特徴抽出装置。

【請求項５】　請求項１あるいは請求項２に記載の話者特徴抽出装置において、

上記ループ学習手段によって最終的にクラスタリングされた各クラスタに属する学習話者を更にクラスタリングしてサブクラスタを生成し、上記各サブクラスタに属する話者集合毎に学習によって１状態の混合ガウス分布型音響モデルを生成するサブクラスタ生成手段を備えて、

上記音響モデル格納部は、上記サブクラスタ生成手段によって生成された１状態の混合ガウス分布型音響モデルの群を、各サブクラスタの伸縮係数αに対応付けて格納するようになっていることを特徴とする話者特徴抽出装置。

【請求項６】　入力話者の音声に基づいて、標準話者の音声と上記入力話者の音声との関係を表わすパラメータを話者特徴として抽出する話者特徴抽出装置において、

上記標準話者に対する声道長の伸縮係数αの値に基づいて学習話者をクラスタリングし、各クラスタに属する話者集合毎に１状態の混合ガウス分布型音響モデルを生成し、この生成された１状態の混合ガウス分布型音響モデルの夫々に対する上記学習話者の音声サンプルの尤度に基づいて上記学習話者を再クラスタリングし、上記１状態の混合ガウス分布型音響モデルの生成と上記学習話者の再クラスタリングとを所定の条件を満たすまで繰り返して最終的に生成された１状態の混合ガウス分布型音響モデルの群が格納された音響モデル格納部と、

上記音響モデル格納部に格納された１状態の混合ガウス分布型音響モデルの夫々に対する入力話者の音声サンプルの尤度を算出し、最大の尤度を呈する１状態の混合ガウス分布型音響モデルを入力話者が属するクラスタの情報として選択する話者クラスタ選択部を備えて、

上記入力話者の特徴として上記最大の尤度を呈する１状態の混合ガウス分布型音響モデルを抽出することを特徴とする話者特徴抽出装置。

3

【請求項７】　音響モデルとして隠れマルコフモデルを用い、入力話者の音声に基づいて抽出された標準話者の音声と上記入力話者の音声との関係を表わすパラメータを話者特徴として、上記入力話者の音声を認識する音声認識装置であって、
請求項１あるいは請求項６に記載の話者特徴抽出装置と、
上記話者特徴抽出装置の音響モデル格納部に格納された各１状態の混合ガウス分布型音響モデルによって表わされる話者クラスタに属する話者集合毎に、学習によって生成された隠れマルコフモデルの群を格納する隠れマルコフモデル格納部と、
上記話者特徴抽出装置によって選択されたクラスタに基づいて、上記隠れマルコフモデル格納部に格納されている上記選択されたクラスタに対応するクラスタの隠れマルコフモデルを、音声認識用の音響モデルとして切り換え選出する切換部を備えたことを特徴とする音声認識装置。

【請求項８】　音響モデルとして隠れマルコフモデルを用い、入力話者の音声に基づいて抽出された標準話者の音声と上記入力話者の音声との関係を表わすパラメータを話者特徴として、上記入力話者の音声を認識する音声認識装置であって、
請求項１あるいは請求項６に記載の音響モデル格納部に格納された各１状態の混合ガウス分布型音響モデルによって表わされる話者クラスタに属する話者集合毎に、学習によって生成された隠れマルコフモデルの群を格納する隠れマルコフモデル格納部と、
上記隠れマルコフモデル格納部に格納された隠れマルコフモデルの夫々に対する入力話者の音声サンプルの尤度を算出し、最大の尤度を呈する隠れマルコフモデルを音声認識用の音響モデルとして切り換え選出する切換部を備えたことを特徴とする音声認識装置。

【請求項９】　声道長の伸縮関数αを用いて入力音声のスペクトルの周波数軸を伸縮することによって入力話者の音響特徴量を標準話者の音響特徴量に正規化する話者正規化手段を有する音声認識装置において、
上記話者正規化手段は、
請求項１あるいは請求項６に記載の話者特徴抽出装置と、
上記入力話者の音声サンプルに基づいて、上記話者特徴抽出装置によって選択された１状態の混合ガウス分布型音響モデルに対応する声道長の伸縮係数αを用いて、上記入力音声のスペクトルの周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴とする音声認識装置。

【請求項１０】　声道長の伸縮関数αを用いて音声のスペクトルの周波数軸を伸縮することによって、音響モデルを入力話者に話者適応させる話者適応手段を有する音声認識装置において、

4

上記話者適応手段は、
請求項１あるいは請求項６に記載の話者特徴抽出装置と、
上記入力話者の音声サンプルに基づいて、上記話者特徴抽出装置によって選択された１状態の混合ガウス分布型音響モデルに対応する声道長の伸縮係数αの逆数を用いて、上記音響モデルのスペクトルの周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴とする音声認識装置。

【請求項１１】　入力話者の音声に基づいて、標準話者の音声と上記入力話者の音声との関係を表わすパラメータを話者特徴として抽出する話者特徴抽出方法であって、
各学習話者に関して、上記標準話者に対する声道長の伸縮係数αを所定の方法によって予め求め、この求められた伸縮係数αの値に基づいて上記学習話者をクラスタリングし、
上記クラスタリングされた各クラスタに属する話者集合毎に、学習によって１状態の混合ガウス分布型音響モデルを生成し、
上記生成された１状態の混合ガウス分布型音響モデルの夫々に対する上記学習話者の音声サンプルの尤度を算出し、その尤度に基づいて上記学習話者を再クラスタリングし、
所定の条件を満たすまで、上記１状態の混合ガウス分布型音響モデルの生成と上記学習話者の再クラスタリングとを繰り返すループ学習を行い、
上記ループ学習によって最終的に生成された１状態の混合ガウス分布型音響モデルの群を音響モデル格納部に格納し、
上記音響モデル格納部に格納された１状態の混合ガウス分布型音響モデルの夫々に対する入力話者の音声サンプルの尤度を算出し、最大の尤度を呈する１状態の混合ガウス分布型音響モデルを上記入力話者の特徴として抽出することを特徴とする話者特徴抽出方法。

【請求項１２】　コンピュータを、
請求項１に記載の学習話者クラスタリング手段，音響モデル生成手段，再クラスタリング手段，ループ学習手段，音響モデル格納部および話者クラスタ選択部として機能させる話者特徴抽出処理プログラムが記録されたことを特徴とするコンピュータ読出し可能なプログラム記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、標準話者の音声スペクトルに対する入力音声スペクトルの周波数軸の線形伸縮係数を話者特徴として抽出する話者特徴抽出装置および話者特徴抽出方法、その抽出方法を用いた音声認識装置，音声合成装置、並びに、話者特徴抽出処理プログラムを記録したプログラム記録媒体に関する。

【０００２】

【従来の技術】従来より、隠れマルコフモデル(Hidden Markov Model：以下、ＨＭＭと言う)を用いた音声認識方法の開発が盛んに行われている。このＨＭＭは、大量の音声データから得られる音声の統計的特徴を確率的にモデル化したものであり、このＨＭＭを用いた音声認識方法の詳細は、中川聖一著「確率モデルによる音声認識」(電子情報通信学会)に詳しい。このＨＭＭに基づく話者照合や話者適応や話者正規化に関する研究が行われている。通常、話者正規化や話者適応技術は、音声データの内容や量に依存するため、少量の発声データからでは安定した性能向上が難しい。そこで、声道長を用いた手法が注目されており、特に声道長に基づく話者正規化が盛んに研究されて効果が出ている。

【０００３】上記声道長は、音声のスペクトルの大まかな特徴を表わすパラメータである。そして、上記声道長の差は話者間の主な変動要因であり、声道長は従来の話者適応法に比べて１個のパラメータあるいは極めて少ないパラメータで音声の特徴を表現できることから、声道長にはより少量の学習データで効率良く正規化できるというメリットがある。

【０００４】ところで、標準話者の音声パターンに対する入力話者の音声サンプルの尤度を最大にするという基準(最尤推定)に従って、上記音声サンプルにおける周波数軸の線形伸縮係数α（声道長正規化係数）を求める（ＭＬ－ＶＴＬＮ法：MaximumLikelihood Vocal Tract Length Normalization)。そして、この声道長伸縮係数αを用いて、入力話者の音声サンプルの周波数軸を線形伸縮して話者正規化を行う技術が提案されている（例えば、AT&T Bell Labs. Li Lee, Richard C. Rose, "Speaker Normalization using Efficient Frequency Warping Procedures", pp. 353-356 ICASSP96 (1996))。また、特開平１１－３２７５９２号公報においては、声道を前室と後室との２つの室に分け、入力音声のフォルマント周波数を用いて、各室に対応した２つの周波数軸線形伸縮係数αを用いて話者正規化する技術が開示されている。

【０００５】尚、上記話者適応は標準となる音響モデルを入力話者に対して適応(つまり正規化)させる技術であり、話者正規化とは表裏一体の関係にある。

【０００６】また、話者クラスタリングを用いた音声認識方法がある。この音声認識方法においては、学習話者間の距離を定義して学習話者をクラスタリングしておき、クラスタ毎にそのクラスタに属する学習話者群の音声データを用いて音響モデルを作成する。そして、認識時には、入力音声に最適なクラスタを選択し、そのクラスタの音響モデルを用いて認識処理を行うのである。その場合における学習話者間の距離として上記声道長の周波数軸線形伸縮係数を用いる音声認識装置が提案されている(特開平１１－１７５０９０号公報)。この公報においては、声道を前室と後室との２つの室に分け、各室に対応した２つの周波数軸線形伸縮係数を用いて学習話者をクラスタリングするようにしている。

【０００７】また、声道長の非線形な伸縮関数を導入してその係数αでクラスタリングする方法や、ＧＭＭ(ガウシアン混合モデル)を用いて話者クラスタリングする方法が提案されている(佐藤他「ＧＭＭによる音響モデル用学習データの自動選択」日本音響学会春季研究発表会講演番号２－８－３　２０００年３月)。上記ＧＭＭは１状態の混合ガウス分布で表現される音響モデルであり、発話内容に因らずに入力音声に声質の近いＧＭＭが大きい値を出力するように設計されている。元々は話者照合における話者モデルとして提案された手法である。

【０００８】

【発明が解決しようとする課題】しかしながら、上記従来の声道長に基づく話者適応や話者正規化には、以下のような問題がある。すなわち、声道長伸縮関数の求め方として学習サンプル全体を対象として最尤推定する方法(ＭＬ－ＶＴＬＮ法)等が提案されている。このような声道長に基づく話者適応や話者正規化は極めて少ないパラメータ数で表現できるとは言うものの、声道長の抽出は発声データの内容や量に大きく左右されるために、少ない学習サンプルから必ずしも安定して声道長を抽出できるとは限らない。したがって、声道長に基づいて話者正規化や話者適応や話者クラスタリングを行う音声認識装置においては、性能劣化を招くと言う問題がある。

【０００９】実際の声道長はＭＲＩ(磁気共鳴画像診断装置)で測定しなければ分からないため、現時点においては直ちに真の声道長を知るのは困難な状況にある。上記特開平１１－３２７５９２号公報および特開平１１－１７５０９０号公報では、声道パラメータを得るために入力音声のフォルマント周波数を用いている。しかしながら、一般的にフォルマント周波数を全自動で求めることは困難であり、上記特開平１１－３２７５９２号公報に開示された線形伸縮係数を用いた話者正規化方法や上記特開平１１－１７５０９０号公報に開示された線形伸縮係数を用いた音声認識装置では、実時間性に欠けるという問題がある。

【００１０】また、上記特開平１１－１７５０９０号公報のごとく、話者クラスタリングを用いた音声認識のアプローチも盛んに試みられているが、大きな性能改善は達成できていない。不特定話者(ＳＩ)音響モデル(すなわち男女共通の音響モデル)をベースラインとすると、男女別(ＧＤ)音響モデルは最もシンプルながら性能向上量が最も大きい。しかしながら、話者クラスタによって更なる細分化(クラスタ化)を行っても効果は薄いという報告がなされており、その場合における単語誤り率(ＷＥＲ：Word Error Rate)の削減は１０％～２０％程度に留まっている。これは、話者間の距離を定義する適当な尺度がないために上手くクラスタリングできなかった

り、クラスタを増やすと１つのクラスタ当りの学習話者数が少なくなってロバスト性に欠けたりするためである。

【００１１】さらに、何れの音響モデルの場合も、各話者クラスタの境界領域では学習サンプルが希薄だったり段差ができたりしているため上手く学習されていない。したがって、入力話者が各クラスタの境界付近に位置する場合には、認識率が劣化するという問題(所謂、hard decision問題)が生ずることになる。尚、個々の学習話者の音響モデル間の距離でクラスタリングを行った場合は、クラスタを木構造にし、入力話者が二つのクラスタの境界付近に位置する場合は上記２つのクラスタの上位ノードのクラスタの音響モデルを採用する方法もある。しかしながら、この方法の場合には、二つのクラスタの境界付近に位置する入力話者に対しては上位ノードの音響モデルを使用するためによりブロードな音響モデルとなってしまい、高い認識率は得にくいのである。

【００１２】ところで、上記ＭＬ－ＶＴＬＮ法に基づいて話者をクラスタリングする場合には、以下のような問題がある。

・真の声道長伸縮係数αの値を求めるのは困難である。上記真の声道長伸縮係数α値を求めるには各話者についてＭＲＩ装置で実測しなければならない。しかしながら、既に構築済みの学習用音声データベースがあり、直ちにはそのデータベースを活用するしかない場合や、上記ＭＲＩ装置を利用し難い環境下にある場合には、音声波形から声道長伸縮係数αの値を自動推定する必要がある。したがって、自動推定する限りにおいてはどうしても推定誤差の問題が付きまとうことになる。

・例え、上記ＭＲＩ装置で測定した実測値をもってしても、発声の仕方の影響があるために、適切な声道長伸縮係数αの値が得られるとは限らない。

【００１３】一方、上記ＧＭＭに基づいて話者をクラスタリングする場合には、一般に以下のような課題がある。

・初期値をランダムにして全自動でクラスタリングし、その後はＨＭＭの学習アルゴリズムに頼っている。しかしながら、音声データは多数の要因が絡み合って複雑な構造を成しているために、このような方法の場合には、音声の微細な特徴を捉えてクラスタリングする危険性がある。

・上記ＧＭＭ間の距離の物理的意味が不明である。つまり、距離の大小が音響的に何に対応しているのかが分からないために、周波数伸縮による話者正規化は適用できない。

・クラスタ化による学習データの減少を補う目的で近傍クラスタの学習データを編入させる場合に上記ＧＭＭ間の距離を用いると、話者の特徴空間上、様々な方向に位置する話者データを編入するになる。その結果、ぼやけた分布になってしまい、精密な話者特徴を抽出できなくなる。したがって、このようにしてできたＧＭＭを基に学習されたＨＭＭに対しても精度の劣化を招くと言う問題がある。

【００１４】以上のごとく、上記話者適応(話者正規化)においては少ない発声データから音響モデルを精度良く適応できないため、誤り率を半減させるためには数十単語以上の発声データが必要となり、学習話者に負担を強いることになるという問題がある。また、音声合成における声質変換の場合にも、同様に少ない発声データからは精度良く声質が得られないという問題がある。

【００１５】そこで、この発明の目的は、より少ない発声データから精度良く話者特徴を抽出できる話者特徴抽出装置および話者特徴抽出方法、その抽出方法を用いた音声認識装置、並びに、話者特徴抽出処理プログラムを記録したプログラム記録媒体を提供することにある。

【００１６】

【課題を解決するための手段】上記目的を達成するため、第１の発明は、入力話者の音声に基づいて，標準話者の音声と上記入力話者の音声との関係を表わすパラメータを話者特徴として抽出する話者特徴抽出装置において、各学習話者に関して，上記標準話者に対する声道長の伸縮係数αを所定の方法によって予め求め，この求められた伸縮係数αの値に基づいて上記学習話者をクラスタリングする学習話者クラスタリング手段と、上記クラスタリングされた各クラスタに属する話者集合毎に，学習によってＧＭＭを生成する音響モデル生成手段と、上記生成されたＧＭＭの夫々に対する上記学習話者の音声サンプルの尤度を算出し，その尤度に基づいて上記学習話者を再クラスタリングする再クラスタリング手段と、上記音響モデル生成手段と再クラスタリング手段とを制御して，所定の条件を満たすまで，上記ＧＭＭの生成と上記学習話者の再クラスタリングとを繰り返すループ学習手段と、上記ループ学習手段によって最終的に生成されたＧＭＭの群を格納する音響モデル格納部と、上記音響モデル格納部に格納されたＧＭＭの夫々に対する入力話者の音声サンプルの尤度を算出し，最大の尤度を呈するＧＭＭを入力話者が属するクラスタの情報として選択する話者クラスタ選択部を備えて、上記入力話者の特徴として上記最大の尤度を呈するＧＭＭを抽出することを特徴としている。

【００１７】上記構成によれば、学習話者をクラスタリングするに際して、先ず、各学習話者を標準話者に対する声道長の伸縮係数αに基づいてクラスタリングし、各クラスタに属する話者集合毎にＧＭＭを生成する。そして，このＧＭＭ群を用いてループ学習を行うことによって学習話者を再クラスタリングするようにしている。こうして、各クラスタの初期値として声道長という大局的な特徴を明示的に与えることによって、各クラスタ間の距離の物理的意味が明確になり、効率よくクラスタリングが行われる。さらに、１状態の混合ガウス分布型音響

モデルであるＧＭＭを用いて学習話者を再クラスタリングすることによって、発話内容に因らずに話者の特徴を良く表わす話者クラスタが得られ、声道長伸縮係数αの抽出誤りも修復されている。

【００１８】したがって、上述のようにして得られた話者クラスタ毎にＧＭＭが格納された音響モデル格納部を用いて、話者クラスタ選択部によって、入力話者の音声サンプルに対して最大尤度を呈するＧＭＭを選択することによって、発話内容に因らずに精度良く入力話者の特徴が抽出される。

【００１９】また、１実施例では、上記第１の発明の話者特徴抽出装置において、上記学習話者クラスタリング手段によってクラスタリングされた各クラスタのうちの注目クラスタに隣接する隣接クラスタに属する学習話者の音声サンプル，または，上記注目クラスタと上記伸縮係数α値の差が所定値以内の近傍クラスタに属する学習話者の音声サンプルに対して，上記注目クラスタと隣接クラスタまたは近傍クラスタとの上記伸縮係数α値に基づいて周波数伸縮を行うことによって上記注目クラスタに属する音声サンプルを生成し，この生成された音声サンプルを上記注目クラスタに編入して当該クラスタの音声サンプル数を豊富化する操作を，上記学習話者クラスタリング手段によってクラスタリングされた総てのクラスタについて実行する音声サンプル豊富化手段を備えて、上記音響モデル生成手段は、上記音声サンプル豊富化手段によって音声サンプル数が豊富化された後の各クラスタ毎に、上記ＧＭＭを生成するようにしている。

【００２０】この実施例によれば、上記学習話者のクラスタリングに際して、ある注目クラスタの隣接クラスタまたは近傍クラスタに属する話者の音声サンプルに対して周波数伸縮が行われ、上記注目クラスタに属する音声サンプルが生成されて注目クラスタに編入される。こうして、学習データ不足が補われて、少ない発声データからでも各クラスタの音響モデルが精密に構築される。

【００２１】また、１実施例では、上記第１の発明の話者特徴抽出装置において、上記注目クラスタに属する音声サンプルを生成する際に、上記音声サンプル豊富化手段が上記学習話者の音声サンプルに対して周波数伸縮を行う音声区間を、有音・無音の別および調音点に基づいて限定するようにしている。

【００２２】この実施例によれば、上記音声サンプル豊富化手段によって、上記隣接クラスタや近傍クラスタの音声サンプルから注目クラスタに属する音声サンプルを生成する際に、上記音声サンプルに対して周波数伸縮を行う音声区間が有音・無音の別および調音点に基づいて限定される。したがって、声道長の差の影響を受け難い音素や無音部を上記周波数軸伸縮の対象外にして、声道長の差の影響を受け難い音素や無音部まで変形されることが防止される。

【００２３】また、１実施例では、上記第１の発明の話者特徴抽出装置において、上記再クラスタリング手段によって上記学習話者を再クラスタリングする場合に、上記再クラスタリングの対象となる対象学習話者が再クラスタリングの前に属していたクラスタの伸縮係数αと再クラスタリング後に属するクラスタの伸縮係数αとが所定値以上離れている場合には、当該対象学習話者を上記再クラスタリングの対象から外すようになっている。

【００２４】この実施例によれば、再クラスタリングの対象となる学習話者が再クラスタリングの前後に属しているクラスタの伸縮係数αが所定値以上離れている場合は、当該対象学習話者を上記再クラスタリングの対象から外すことによって、声道長伸縮係数αが極端に異なる話者同士が同じクラスタに属することが防止される。

【００２５】また、１実施例では、上記第１の発明の話者特徴抽出装置において、上記ループ学習手段によって最終的にクラスタリングされた各クラスタに属する学習話者を更にクラスタリングしてサブクラスタを生成し，上記各サブクラスタに属する話者集合毎に学習によってＧＭＭを生成するサブクラスタ生成手段を備えて、上記音響モデル格納部は、上記サブクラスタ生成手段によって生成されたＧＭＭの群を、各サブクラスタの伸縮係数αに対応付けて格納するようになっている。

【００２６】この実施例によれば、上記ループ学習手段によって最終的にクラスタリングされた各クラスタがさらにクラスタリングされてサブクラスタが生成される。このサブクラスタは声道長以外の要因にも対応することができ、より精密な話者特徴が抽出される。

【００２７】また、第２の発明は、入力話者の音声に基づいて標準話者の音声と上記入力話者の音声との関係を表わすパラメータを話者特徴として抽出する話者特徴抽出装置において、上記標準話者に対する声道長の伸縮係数αの値に基づいて学習話者をクラスタリングし，各クラスタに属する話者集合毎にＧＭＭを生成し，この生成されたＧＭＭ夫々に対する上記学習話者の音声サンプルの尤度に基づいて上記学習話者を再クラスタリングし，上記ＧＭＭの生成と上記学習話者の再クラスタリングとを所定の条件を満たすまで繰り返して最終的に生成されたＧＭＭの群が格納された音響モデル格納部と、上記音響モデル格納部に格納されたＧＭＭの夫々に対する入力話者の音声サンプルの尤度を算出し，最大の尤度を呈するＧＭＭを入力話者が属するクラスタの情報として選択する話者クラスタ選択部を備えて、上記入力話者の特徴として上記最大の尤度を呈するＧＭＭを抽出することを特徴としている。

【００２８】上記構成によれば、標準話者に対する声道長の伸縮係数αに基づいて学習話者をクラスタリングし、各クラスタ毎のＧＭＭの生成とそのＧＭＭ群を用いた学習話者の再クラスタリングとを所定の条件を満たすまで繰り返し、最終的に生成されたＧＭＭ群を話者クラスタ毎に格納した音響モデル格納部を用いて、話者クラ

スタ選択部によって、入力話者の音声サンプルに対して最大尤度を呈するＧＭＭが選択される。こうして、発話内容に拘らず精度良く入力話者の特徴が抽出される。

【００２９】また、第３の発明は、音響モデルとしてＨＭＭを用い，入力話者の音声に基づいて抽出された標準話者の音声と上記入力話者の音声との関係を表わすパラメータを話者特徴として上記入力話者の音声を認識する音声認識装置であって、上記第１の発明または第２の発明の話者特徴抽出装置と、上記話者特徴抽出装置の音響モデル格納部に格納された各ＧＭＭによって表わされる話者クラスタに属する話者集合毎に，学習によって生成されたＨＭＭの群を格納するＨＭＭ格納部と、上記話者特徴抽出装置によって選択されたクラスタに基づいて，上記ＨＭＭ格納部に格納されている上記選択されたクラスタに対応するクラスタのＨＭＭを音声認識用の音響モデルとして切り換え選出する切換部を備えたことを特徴としている。

【００３０】上記構成によれば、上記第１の発明または第２の発明の話者特徴抽出装置の音響モデル格納部における話者クラスタ毎に生成されたＨＭＭの群が格納されたＨＭＭ格納部から、切換部によって、上記話者特徴抽出装置で選択された話者クラスタのＨＭＭが音声認識用の音響モデルとして切り換え選出される。こうして、発話内容に拘らずに少ない音声データで、上記入力話者の特徴を精度良く表わすＨＭＭを用いて、入力話者の音声が正確に認識される。

【００３１】また、第４の発明は、音響モデルとしてＨＭＭを用い，入力話者の音声に基づいて抽出された標準話者の音声と上記入力話者の音声との関係を表すパラメータを話者特徴として上記入力話者の音声を認識する音声認識装置であって、上記第１の発明あるいは第２の発明に係る音響モデル格納部に格納された各ＧＭＭによって表わされる話者クラスタに属する話者集合毎に，学習によって生成されたＨＭＭの群を格納するＨＭＭ格納部と、上記ＨＭＭ格納部に格納されたＨＭＭの夫々に対する入力話者の音声サンプルの尤度を算出し最大の尤度を呈するＨＭＭを音声認識用の音響モデルとして切り換え選出する切換部を備えたことを特徴としている。

【００３２】上記構成によれば、上記第１の発明または第２の発明に係る音響モデル格納部の話者クラスタ毎に生成されたＨＭＭの群が格納されたＨＭＭ格納部を用いて、切換部によって、最大の尤度を呈するＨＭＭが音声認識用の音響モデルとして切り換え選出される。こうして、発話内容に拘らずに少ない音声データで、上記入力話者の特徴を精度良く表わすＨＭＭを用いて、入力話者の音声が正確に認識される。

【００３３】また、第５の発明は、声道長の伸縮関数αを用いて入力音声のスペクトルの周波数軸を伸縮することによって入力話者の音響特徴量を標準話者の音響特徴量に正規化する話者正規化手段を有する音声認識装置において、上記話者正規化手段は、上記第１の発明あるいは第２の発明の話者特徴抽出装置と、上記入力話者の音声サンプルに基づいて上記話者特徴抽出装置によって選択されたＧＭＭに対応する声道長の伸縮係数αを用いて，上記入力音声のスペクトルの周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴としている。

【００３４】上記構成によれば、入力話者の音声サンプルに基づいて上記第１の発明あるいは第２の発明の話者特徴抽出装置によって選択されたＧＭＭに対応する声道長の伸縮係数αを用いて、周波数ワープ手段によって上記入力音声のスペクトルの周波数軸が伸縮されて、上記入力話者の音響特徴量が標準話者の音響特徴量に正規化される。こうして、発話内容に拘らずに少ない音声データで、上記入力話者の特徴を精度良く表わす声道長の伸縮係数αを用いて、より標準話者の音響特徴量に近づくように話者正規化が行われる。その結果、高い音声認識率が得られる。

【００３５】また、第６の発明は、声道長の伸縮関数αを用いて音声のスペクトルの周波数軸を伸縮することによって音響モデルを入力話者に話者適応させる話者適応手段を有する音声認識装置において、上記話者適応手段は、上記第１の発明あるいは第２の発明の話者特徴抽出装置と、上記入力話者の音声サンプルに基づいて上記話者特徴抽出装置によって選択されたＧＭＭに対応する声道長の伸縮係数αの逆数を用いて，上記音響モデルのスペクトルの周波数軸を伸縮する周波数ワープ手段で構成されていることを特徴としている。

【００３６】上記構成によれば、入力話者の音声サンプルに基づいて、上記第１の発明あるいは第２の発明の話者特徴抽出装置によって選択されたＧＭＭに対応する声道長の伸縮係数αの逆数を用いて、周波数ワープ手段によって音響モデルのスペクトルの周波数軸が伸縮されて上記音響モデルが入力話者に話者適応される。こうして、発話内容に拘らずに少ない音声データで、上記入力話者の特徴を精度良く表わす声道長の伸縮係数αの逆数を用いて、より入力話者の音響特徴量に近づくように話者適応が行われる。その結果、高い音声認識率が得られる。

【００３７】また、第７の発明は、入力話者の音声に基づいて，標準話者の音声と上記入力話者の音声との関係を表すパラメータを話者特徴として抽出する話者特徴抽出方法であって、各学習話者に関して，上記標準話者に対する声道長の伸縮係数αを所定の方法によって予め求め，この求められた伸縮係数αの値に基づいて上記学習話者をクラスタリングし、上記クラスタリングされた各クラスタに属する話者集合毎に学習によってＧＭＭを生成し、上記生成されたＧＭＭの夫々に対する上記学習話者の音声サンプルの尤度を算出し，その尤度に基づいて上記学習話者を再クラスタリングし、所定の条件を満た

すまで上記ＧＭＭの生成と上記学習話者の再クラスタリングとを繰り返すループ学習を行い、上記ループ学習によって最終的に生成されたＧＭＭの群を音響モデル格納部に格納し、上記音響モデル格納部に格納されたＧＭＭの夫々に対する入力話者の音声サンプルの尤度を算出し，最大の尤度を呈するＧＭＭを上記入力話者の特徴として抽出することを特徴としている。

【００３８】上記構成によれば、上記第１の発明の場合と同様に、学習話者をクラスタリングする際に、各クラスタの初期値として声道長という大局的な特徴を明示的に与えることによって、各クラスタ間の距離の物理的意味が明確になり、効率よくクラスタリングが行われる。さらに、１状態の混合ガウス分布型音響モデルであるＧＭＭを用いて学習話者を再クラスタリングすることによって、発話内容に因らずに話者の特徴を良く表わすクラスタが得られ、声道長伸縮係数αの抽出誤りも修復されている。

【００３９】したがって、上述のようにして得られた話者クラスタ毎にＧＭＭが格納された音響モデル格納部を用いて、入力話者の音声サンプルに対して最大尤度を呈するＧＭＭを選択することによって、発話内容に因らずに精度良く入力話者の特徴が抽出される。

【００４０】また、第８の発明のプログラム記録媒体は、コンピュータを、上記第１の発明に係る学習話者クラスタリング手段，音響モデル生成手段，再クラスタリング手段，ループ学習手段，音響モデル格納部および話者クラスタ選択部として機能させる話者特徴抽出処理プログラムが記録されていることを特徴としている。

【００４１】上記構成によれば、上記第１の発明の場合と同様に、発話内容に因らずに話者の特徴を良く表わす話者クラスタ毎にＧＭＭが格納された音響モデル格納部を用いて、入力話者の音声サンプルに対して最大尤度を呈するＧＭＭを選択することによって、発話内容に因らずに精度良く入力話者の特徴が抽出される。

【００４２】

【発明の実施の形態】以下、この発明を図示の実施の形態により詳細に説明する。

＜第１実施の形態＞図１は、本実施の形態の音声認識装置におけるブロック図である。尚、この音声認識装置は、話者クラスタリング方式を用いた音声認識装置である。音声入力部１において、マイクから入力された音声はディジタル波形に変換されて音響分析部２に入力される。音響分析部２は、入力されたディジタル波形を短い時間間隔（フレーム）毎に周波数分析し、スペクトルを表す音響パラメータのベクトル系列に変換する。ここで、上記周波数分析には、ＭＦＣＣ（メル周波数ＦＦＴ（高速フーリエ変換）ケプストラム）やＬＰＣ（線形予測分析）メルケプストラム等のスペクトルを効率よく表現できる方法が用いられる。こうして得られた音響パラメータ系列は、話者クラスタ選択部３及び尤度（音韻類似度）演算部４に送出される。

【００４３】上記話者クラスタ選択部３は、ＧＭＭ格納部６と共に話者特徴抽出部１１を構成し、以下のようにして話者特徴としてのクラスタ情報を生成する。すなわち、話者クラスタ選択部３は、入力された音響パラメータ系列にＧＭＭ格納部６に話者クラスタ別に格納されたｎ個のＧＭＭの夫々を作用させて尤度を算出する。そして、算出されたｎ個の尤度のうちの最大値を与えるＧＭＭのインデックス（ｉ）（ｉ＝１，２，…，ｎ）を、その入力話者に適合した話者クラスタ情報として出力する。ここで、上記ＧＭＭは、１状態からなる混合ガウス分布で表現される。

【００４４】切換部５は、話者クラスタ音響モデル格納部７に話者クラスタ別に格納された音響モデル（本実施の形態ではＨＭＭを使用）の中から、話者クラスタ選択部３からのクラスタ情報に適合する話者クラスタの音響モデルを切り換え選択して尤度演算部４に送出する。そうすると、尤度演算部４は、音響分析部２からの入力音声の音響パラメータベクトルに対して切換部５からの音響モデルを作用させて、各音韻の状態毎に尤度を算出する。そして、得られた尤度系列を照合部８に送出する。

【００４５】上記照合部８は、上記尤度演算部４からの尤度系列に対して、言語モデル格納部９に登録された総ての言語モデル（単語）との照合を行ない、各単語のスコアを算出する。そして、上位のスコアを呈する単語を認識候補（認識結果）として出力部１０から出力するのである。

【００４６】ここで、上記話者クラスタ選択の方法には、以下の[ａ]および[ｂ]に示す２通りの方法がある。本実施の形態においては[ａ]の方法を用いている。

[ａ] ＧＭＭ格納部６の利用

[ｂ] 話者クラスタ音響モデル格納部７の利用

【００４７】上記[ａ]の方法は、各話者クラスタに対して１つのＧＭＭを作成しておく。そして、入力音声に対して各ＧＭＭを作用させてＧＭＭ毎の尤度を算出し、最も大きい尤度を算出したＧＭＭに対応する話者クラスタを選択するのである。その場合、入力音声の正解音素列をユーザが教える必要がなく、教師なしでクラスタが選択できる。すなわち、エンロールモードがないシステムにおいて有効なのである。

【００４８】また、上記[ｂ]の方法は、上記話者クラスタの音響モデル自身の尤度を用いる方法である。入力音声に対して教師語彙が与えられ、各話者クラスタにおける教師語彙の音響モデルを用いて認識処理を行って、話者クラスタ毎の尤度を算出する。そして、最も大きい尤度を呈する話者クラスタを選択するのである。この選択方法は、エンロールモードにおいて入力音声の正解音素列をユーザが教えるという教師あり学習を基本としている。認識処理と同じ高精度な音響モデルを用いるので計算量は多くなるが、エンロールによって正確なクラスタ

選択が可能となる。

【００４９】ここで、本実施の形態における話者クラスタ音響モデル格納部７は、学習話者のクラスタ数ｎに応じて、第１音響モデル格納部７ａ,第２音響モデル格納部７ｂ,…,第ｎ音響モデル格納部７ｎのｎ個の音響モデル格納部で構成されている。ここで、各音響モデル格納部７ａ～７ｎに格納される各音響モデルは、混合ガウス分布型のＨＭＭである。この発明においては、生理的な特徴の変動に対処可能にすることを目的としており、話者性の大局的な安定要因である声道長を初期値としたＧＭＭをクラスタリング対象にするのである。

【００５０】尚、生理的な特徴の変動要因としては、上記声道長以外にも鼻腔,副鼻腔,声帯等の多くの要因があり、それらが絡み合って複雑な特徴を成している。したがって、個々の要因を数理的に扱うのは得策ではない。そこで、本実施の形態においては、各要因の複雑な特徴を混合ガウス分布型ＨＭＭ(音響モデル)や、同一クラスタ内での複数のＧＭＭによるサブクラスタで表現するのである。

【００５１】以下、上記ＧＭＭ格納部６に格納されるＧＭＭ群の作成方法について詳細に説明する。本実施の形態におけるＧＭＭ群の作成方法は、下記の処理手順によって行われる。これらの処理はＧＭＭ学習部１２によって、学習データ格納部１３に格納された学習話者の音声データを用いて行われる。尚、記憶領域や処理量が膨大になるために、予めオフライン処理によって作成されている。作成されたＧＭＭ群は、通常はＲＯＭ(リード・オンリ・メモリ)やフラッシュメモリやハードディスク等で構成されるＧＭＭ格納部６に格納される。尚、学習データ格納部１３に格納された全学習話者の集合を男女別に２分割してもよい。その湯合には、男性用話者クラスタと女性用話者クラスタとの合計２種類のＧＭＭ群が生成されることになる。

【００５２】① 夫々の学習話者の音声データに所定の方法によって声道長伸縮係数αを与え、声道長伸縮係数αの値に応じて学習話者をＮ個にクラスタリングしておく。尚、ＧＭＭの初期モデルの混合数はＭとする。

② あるクラスタＣに隣接するクラスタに属する音声データ、または、あるクラスタＣとの声道長伸縮係数αの値の差が所定値δ以内であるクラスタに属する音声データに対して、ある特定の区間を対象とした周波数伸縮を行うことによって上記クラスタＣに属する音声データを生成する。そして、この生成された音声データをクラスタＣに編入する。こうして、クラスタＣの音声データを豊富化するのである。この豊富化処理を総てのクラスタについて行う。

③ ＧＭＭパラメータを各クラスタ内の音声データからＭＬ(最尤)アルゴリズムにより推定することによって、各クラスタのＧＭＭを作成する。

④ 作成されたＧＭＭに対する各学習話者の音声データのフレーム平均尤度を算出する。

⑤ ある話者の音声データに対して最も高いフレーム平均尤度を与えるＧＭＭのクラスタに、その話者を移動させる(編入する)。但し、そのクラスタの声道長伸縮係数α値に比べて、声道長伸縮係数α値が所定値ε以上離れている話者については、そのクラスタには編入しない。

⑥ 移動させる話者がなくなるか、予め設定した最大の繰り返し回数になるまで③～⑤の処理を繰り返す。

⑦ 混合数を１つ増加して③～⑥の処理を行う。

⑧ 所望の混合数になるまで③～⑦の処理を繰り返す。

⑨ ①～⑧で得られた学習話者のクラスタ結果を基に、サブクラスタ化したＧＭＭを作成する(オプション)。

【００５３】尚、上記ＧＭＭ群の作成処理手順①における各学習話者に対する声道長伸縮係数αの付与は、ＭＬ－ＶＴＬＮ法や、広母音の第２フォルマントの存在領域以下の領域を部分的に補正した非線形周波数ワーピング関数を用いたＭＬ法等によって与えられる。尚、声道長伸縮係数αの推定に際しては、後述する音声区間の分類を用いてもよい。また、ＭＲＩ装置を利用できる等、各学習話者の声道長を実測できる環境にある場合には、実測されたαを用いてもよい。さらに、クラスタ数Ｎと混合数の初期値Ｍとは、学習データ量や声道長伸縮係数αの信頼性に依存するが、例えばＮ＝１２,Ｍ＝２０等とする。Ｎ＝１２とは、α軸上の区間(０.８８，１.１２)を０.０２刻みに分割することに相当する。また、各クラスタの境界をオーバーラップするような分割を許してもよい。

【００５４】また、上記ＧＭＭ群の作成処理手順②における音声データの変換の際には、線形周波数伸縮関数や、広母音の第２フォルマントの存在領域以下の領域を部分的に補正した折れ線周波数伸縮関数を用いる。以下、②の処理内容について詳しく述べる。あるクラスタＣに隣接するクラスタ、または、あるクラスタＣとの声道長伸縮係数αの値の差が所定値δ以内であるクラスタに属する話者の音声データに対して、声道長伸縮係数α値に基づいて周波数伸縮を行うことによって、クラスタＣに属する音声データを生成するのである。例えば、α＝１.０５のクラスタＤに属する音声データに基づいてα＝１.０３のクラスタＣに属する音声データを生成する際には、周波数を０.９８だけ伸縮する。そして、生成された学習話者の音声データをクラスタＣに編入するのである。尚、生成前の音声データは元のクラスタＤに属したままにしておく。但し、学習話者数および１話者当りのデータ量が多量にある場合は、この処理は省略しても差し支えない。逆に、学習話者数や１話者当りのデータ量が少ない場合には、上記所定値δを大きめ(例えばδ＝０.０５)に設定して編入させるデータを増やす。

【００５５】上記ＧＭＭ群の作成処理手順②における音声データの変換の際に、周波数伸縮を行う対象となる音声区間の分類については後述する。尚、伸縮量が非常に

17

小さい場合には全区間を周波数伸縮対象としてもよい。

【００５６】上記ＧＭＭ群の作成処理手順⑤において、上記所定値εの値は、①で与えられた声道長伸縮係数αの信頼度に依存する。上記ＭＲＩ装置によって実測した場合のように声道長伸縮係数αの信頼度が高い場合は移動を禁止してもよいし、所定値εを小さい値に設定してもよい(例えばε＝０．０２)。逆に、声道長伸縮係数αの信頼度が低い場合には、所定値εの値を大きくする(例えばε＝０．０４)。ところで、上記移動の際に、通常は周波数伸縮(αの値を書き換えることに相当)を行わない。但し、所定値εの値を大きめに設定した場合には周波数伸縮を行ってもよい。

【００５７】上記ＧＭＭ群の作成処理手順⑨におけるサブクラスタ化はオプションであり、処理方法は後述する。

【００５８】次に、上記ＧＭＭ群の作成処理手順①にお＊

18

＊ける声道長伸縮係数αの推定や、②における周波数伸縮の際に、対象とする音声区間の分類について説明する。上述の例において、クラスタＤに属している音声データをクラスクＣに変換すると言うことはクラスタＣに正規化していることを意味しているので、ここでは、②の周波数伸縮を単に正規化と呼ぶことにする。先ず、入力話者の音声データに、不特定話者用音響モデルまたは選択された話者クラスタ音響モデルまたは特定話者音響モデルから選択されたものを用いたビタビアルゴリズムによって、音素境界情報を求めておく。

【００５９】次に、その境界情報に基づいて、上記音声データのうち推定/正規化処理の対象となる区間を制御するのである。本実施の形態においては、この推定/正規化処理の対象となる区間を、表１に示す５種類に分類する。

表１

分類	音素	説明
[a]	無音区間	環境騒音そのものであり、声道長とは全く関係ない。
[b]	調音点が歯茎より前に位置する子音	音源が声道出口付近に位置するために、声道全体による共鳴管が形成されず、声道長の影響を受け難い。
[c]	調音点が歯茎より後に位置する子音，半母音	母音と同様に声道長に直接影響される。しかしながら、子音や半母音区間は元来安定していないので声道長の推定には適さない。
[d]	「ウ」を除く母音	母音は全般的に声道長の影響を直接受ける。
[e]	母音「ウ」,撥音	日本語の「ウ」は発声の仕方によってフォルマンと周波数が大きく変動するので、声道長の影響よりも発声の仕方に大きく影響される。撥音も音素環境に大きく依存すると共に鼻音化の影響が大きい。

【００６０】そして、この分類に基づいて、以下のような区別に従って、推定/正規化時におけるＧＭＭ学習部１２の処理を制御するのである。

・上記ＧＭＭ群の作成処理手順①における
声道長伸縮係数αの初期値推定時…分類[d]
・上記ＧＭＭ群の作成処理手順②における
正規化時…分類[c],分類[d],分類[e],(分類[b])

但し、上記正規化時には分類[b]を含めてもよい。発音の仕方によっては、音素「イ」も音素「ウ」と同様に狭母音なのでフォルマント周波数が大きく変動する場合がある。したがって、分類[e]に音素「イ」を含め、分類[d]から音素「イ」を除いてもよい。

【００６１】次に、上記ＧＭＭ群の作成処理手順⑨におけるサブクラスタの作成方法について説明する。サブクラスタの作成は、上記ＧＭＭ群の作成処理手順①～⑧によって得られた総てのクラスタに対して、下記の処理手順を繰り返すことによって行う。

【００６２】Ａ）クラスタ内の学習話者をランダムにＬ個に分割する。すなわち、Ｌ個のサブクラスタを与えるのである。但し、ＧＭＭ初期モデルの混合数は１とする。

Ｂ）ＧＭＭパラメータを各サブクラスタ内の音声データからＭＬアルゴリズムにより推定することによって、各サブクラスタのＧＭＭを作成する。

Ｃ）作成されたＧＭＭに対する各学習話者の音声データのフレーム平均尤度を算出する。

Ｄ）ある話者の音声データに対して最も高いフレーム平均尤度を与えるＧＭＭのサブクラスタに、その話者を移動させる。

Ｅ）移動させる話者がなくなるか、予め設定した最大の繰り返し回数になるまでＢ)～Ｄ)の処理を繰り返す。

Ｆ）混合数を１つ増加してＢ)～Ｅ)の処理を行う。

Ｇ）所望の混合数になるまでＢ)～Ｆ)の処理を繰り返す。

【００６３】上記サブクラスタの作成処理手順Ａ)におけるサブクラスタ数Ｌの値は、クラスタ内の学習話者数及び１話者当りのデータ量に依存するが、通常２～１０の間に設定する。クラスタによってサブクラスタ数Ｌの値を変えてもよい。例えば、分布の中心であるα＝１．０付近のクラスタは話者数が多いためＬ＝５とする。一方、分布の周辺であるα＝０．９付近やα＝１．１付近のクラスタは話者数が少ないためＬ＝２等にするのである。また、上記サブクラスタの作成処理手順Ｇ)における所望の混合数も、サブクラスタ内の学習話者数および１話者当りのデータ量に依存させてもよい。

【００６４】尚、本実施の形態における話者クラスタリ

ングの場合や第２実施の形態における話者正規化の場合でのαは、入力音声から標準音声への正規化係数である。これに対して、第３実施の形態における話者適応の場合でのαは、標準音声から入力音声への写像係数である。このように、両者は裏表の関係であるため、αの値は逆数の関係になる。

【００６５】次に、上記話者クラスタ音響モデル格納部７に格納されるＨＭＭ群の作成方法について説明する。上記ＧＭＭ作成時においてクラスタリングされた話者クラスタ毎にＭＬアルゴリズムを用いて学習することによって、混合ガウス分布型のＨＭＭが作成される。尚、上記ＧＭＭ群の作成処理手順⑨においてサブクラスタ化が行われている場合には、サブクラスタ毎に、同様の手法によって混合ガウス分布型のＨＭＭが作成される。これらの処理はＨＭＭ学習部１４で行われるのであるが、記憶領域や処理量が膨大になるため予めオフライン処理によって作成されている。そして、作成されたＨＭＭ群は、通常はＲＯＭやフラッシュメモリやハードディスク等で構成される話者クラスタ音響モデル格納部７に格納される。

【００６６】上記構成において、入力音声の認識時には以下のように動作する。先ず、話者クラスタ選択部３によって、上述のようにして最適な話者クラスタが選択され、クラスタ情報(ｉ)が切換部５に送出される。次に、尤度演算部４によって、上記切換部５で切り換え選択された話者クラスタの音響モデルを用いて尤度演算が行われ、得られた尤度系列が照合部８に送出される。そして、照合部８によって、ビタビサーチ等の探索アルゴリズムが用いられて言語モデル格納部９の言語モデルとの照合が行われ、各単語のスコアが算出される。尚、本実施の形態においては、照合部８による照合処理の前段処理が訴求点であるから、照合部８に関する詳細な説明は省略する。

【００６７】上述したように、上記ＧＭＭは１状態の混合ガウス分布で表現される音響モデルであって、発話内容に因らずに入力音声に声質の近いＧＭＭが大きい値を出力するように設計されており、話者の特徴を表わすには好適である。ところが、ＧＭＭ間の距離の物理的意味が不明であるため、ＧＭＭ間の距離でクラスタ化を行った場合にはぼやけた分布になってしまい、精密な話者特徴を抽出できないという問題がある。

【００６８】そこで、本実施の形態においては、ＧＭＭ学習部１２で学習話者をクラスタリングするに際して、先ず、学習データ格納部１３に格納された各学習話者の音声データに、声道長正規化手法によって求められた声道長伸縮(正規化)係数αの値を与える。そして、この声道長伸縮係数αに基づいて学習話者をクラスタリングし、各クラスタに属する話者集合毎に所定の学習方法によって１状態の混合ガウス分布型音響モデルであるＧＭＭを生成する。そして、次に、このＧＭＭ群を用いてループ学習を行って、学習話者をクラスタリングし直すのである。

【００６９】音声データは多数な要因がからみあって複雑な構造を成しているので、ＧＭＭ間の距離でクラスタ化を行った場合には音声の微細な特徴を捉えてクラスタリングしてしまう危険性がある。そこで、上述のように、声道長という大局的な特徴を明示的に与えることによって、距離の物理的意味が明確になって、以後の学習をスムーズに実行でき、より効率よくクラスタリングできるのである。さらに、各クラスタのＧＭＭを生成し、このＧＭＭ群を用いて学習話者をクラスタリングし直すようにしている。したがって、発話内容に因らずに話者の特徴を良く表わすクラスタを得ることができ、声道長伸縮係数αの抽出誤りも修復できるのである。

【００７０】また、上記学習話者のクラスタリングに際して、あるクラスタＣとの声道長伸縮係数αの差がδ以内である近傍のクラスタＤに属する話者の音声データに基づいて、声道長正規化手法によって上記クラスタＣに属する音声データを生成するようにしている。したがって、話者をクラスタリングする際における学習データ不足を補うことができ、各話者クラスタの音響モデルを精密に構築できるのである。このことは、音声認識時におけるクラスタの選択(声道長の推定)をより正確に且つ安定して行うことができ、認識率の向上につながる。また、分布の周辺に位置する話者の認識率を向上させることができる。

【００７１】また、上述のようにして上記声道長に基づいて求められた１つの話者クラスタを複数のサブクラスタに分割し、夫々のサブクラスタのＧＭＭを生成してこれを話者特徴とする。こうして、生成されたサブクラスタは声道長以外の要因にも対応することができるので、上記サブクラスタを用いることによってより精密な話者特徴を抽出することができるのである。

【００７２】したがって、上述のようにしてクラスタリングされた話者クラスタ毎に作成されたＨＭＭ群を用いて入力音声を認識することによって、高い認識率を得ることができるのである。

【００７３】尚、上記実施の形態においては、上記話者クラスタ選択部３によって最適な話者クラスタを一つ選択するようにしているが、最適な話者クラスタを含む上位複数の話者クラスタを選択するようにしてもよい。例えば、尤度の上位からｋ個の話者クラスタを選択するとする。そうすると、切換部５によって切り換え選択されたｋ個の音響モデルの夫々に関して、尤度演算部４によって尤度演算が行われて、照合部８にｋ個の尤度系列が送られることになる。したがって、照合部８では、夫々の尤度系列に関して照合処理が行なわれ、最も大きい尤度を与する単語/単語列が認識結果となるのである。

【００７４】また、音声認識装置のハードウェア規模が大きく、計算量が許すのであれば、話者クラスタ選択部

３による話者クラスタ選択を行なわず、尤度演算部において総ての話者クラスタの音響モデルを用いて尤度演算処理を実行するようにしてもよい。この場合、各音響モデルを適用して得られた尤度が最大値を呈する単語/単語列が認識結果となる。

【００７５】＜第２実施の形態＞図２は、本実施の形態の音声認識装置におけるブロック図である。尚、この音声認識装置は、話者正規化方式を用いた音声認識装置であり、話者正規化部２６を有している。音声入力部２１，音響分析部２２，尤度演算部２４，照合部２８，言語モデル格納部２９および出力部３０は、図１に示す上記第１実施の形態における音声入力部１，音響分析部２，尤度演算部４，照合部８，辞書格納部９および出力部１０と同様である。

【００７６】上記話者正規化部２６は、話者特徴抽出部２５と周波数ワープ部２３とから構成される。話者特徴抽出部２５は、図１に示す上記第１実施の形態における話者特徴抽出部１１と同様であり、入力された音響パラメータ系列に対して最大値を与えるＧＭＭのクラスタ情報をＧＭＭ格納部(図示せず)から抽出して話者特徴とする。そして、得られたクラスタ情報から周波数伸縮係数αを得、周波数ワープ部２３に送出する。

【００７７】そうすると、上記周波数ワープ部２３は、この周波数伸縮係数αを係数とする線形周波数ワーピング関数を用いて、入力音声の音声パラメータ系列を周波数ワープ(話者正規化)し、周波数ワープ後の音響パラメータ系列を尤度演算部２４に送出するのである。そして、上記尤度演算部２４では、周波数ワープされた音響パラメータ系列に対して、正規化不特定話者音響モデル格納部２７に格納された不特定話者モデル(ＨＭＭ)を作用させて、各音韻の状態毎に尤度を算出するのである。

【００７８】ここで、上記正規化不特定話者音響モデル格納部２７に格納される不特定話者モデルは、総ての学習話者を周波数伸縮によってα＝１となるように正規化してから通常のＨＭＭ学習を行って作成される。尚、学習話者が多量に存在する場合には、全学習話者のうち、α＝１の話者およびその周辺の話者を正規化して学習の対象としてもよい。

【００７９】上記第１実施の形態におけるＧＭＭ群作成時における話者クラスタリングの場合と同様に、音声認識時における話者正規化部２６による話者正規化と、正規化不特定話者音響モデル格納部２７に格納される不特定話者モデルの学習との場合にも、表１に示す推定/正規化処理の対象となる音素の分類に従って、以下のように正規化対象とする音声区間を限定してもよい。

・音声認識時における話者正規化時…[ｃ],[ｄ],[ｅ],([ｂ])

・不特定話者モデルの学習時…[ｃ],[ｄ],[ｅ],([ｂ])

【００８０】上述のように、本実施の形態においては、上記話者正規化部２６によって入力話者を正規化する際に、話者特徴抽出部２５によって、上記第１実施の形態における話者特徴抽出部１１の場合と同様にして、ＧＭＭ格納部に各話者クラスタ毎に格納されたＧＭＭを入力音響パラメータ系列に作用させて、最大尤度を与えるＧＭＭのインデックス(係数α)をクラスタ情報として求める。そして、周波数ワープ部２３によって、上記クラスタ情報(係数α)を係数とする線形周波数ワーピング関数を用いて入力音声の音響パラメータ系列を周波数ワープすることによって、話者正規化するようにしている。

【００８１】その場合、上記話者特徴抽出部２５が用いるＧＭＭ格納部には、上記第１実施の形態におけるＧＭＭ格納部６の場合と同様に、各学習話者の音声データを声道長伸縮係数αに基づいてクラスタリングし、あるクラスタＣの近傍のクラスタＤに属する話者の音声データに基づいて上記クラスタＣに属する音声データを生成し、各クラスタのＧＭＭを用いて学習話者をクラスタリングし直したものが格納されている。したがって、話者特徴抽出部２５は、少ない学習データから、発話内容に因らずに話者の特徴を良く表わすクラスタ情報を得ることができる。その結果、高い認識率を得ることができるのである。

【００８２】＜第３実施の形態＞図３は、本実施の形態の音声認識装置におけるブロック図である。尚、この音声認識装置は、話者適応方式を用いた音声認識装置であり、話者適応部３６を有している。音声入力部３１，音響分析部３２，尤度演算部３７，照合部３９，言語モデル格納部４０および出力部４１は、図１に示す上記第１実施の形態における音声入力部１，音響分析部２，尤度演算部４，照合部８，言語モデル格納部９および出力部１０と同様である。また、話者特徴抽出部３３は、図２に示す上記第２実施の形態における話者特徴抽出部２５と同様である。

【００８３】上記話者適応部３６は、上記話者特徴抽出部３３と周波数ワープ部３４とから構成される。話者特徴抽出部３３は、上記第２実施の形態の場合と同様にして、入力された音響パラメータ系列に対して最大値を与えるＧＭＭのクラスタ情報をＧＭＭ格納部(図示せず)から抽出して話者特徴とする。そして、得られたクラスタ情報から周波数伸縮係数αを得、周波数ワープ部３４に送出する。

【００８４】そうすると、上記周波数ワープ部３４は、この得られた周波数伸縮係数αの逆数を係数とする線形周波数ワーピング関数を用いて、正規化不特定話者音響モデル格納部３５に格納された不特定話者モデルを周波数ワープする。その場合の周波数ワープに際しては、上記第１実施の形態におけるＧＭＭ群作成時における話者クラスタリングの場合と同様に、表１に示す正規化処理の対象となる音素の分類に従って、以下のように適応化対象とする音声区間を限定するのである。

・音声認識時における周波数ワープ時…[ｂ],[ｃ],

[d],[e]
但し、声道長の影響を受け難い[b]は変換しなくてもよい。

【００８５】こうして周波数ワープされた不特定話者音響モデルは、話者適応モデル(ＨＭＭ)として話者適応音響モデル格納部３８に格納される。そうすると、尤度演算部３７は、音響分析部３２からの入力音声の音響パラメータ系列に対して、話者適応音響モデル格納部３８に格納された話者適応モデルを作用させて、上述した尤度演算処理を行なうのである。

【００８６】その場合、上記話者特徴抽出部３３が用いるＧＭＭ格納部には、上記第１実施の形態におけるＧＭＭ格納部６の場合と同様に、各学習話者の音声データを声道長伸縮係数αに基づいてクラスタリングし、あるクラスタＣの近傍のクラスタＤに属する話者の音声データに基づいて上記クラスタＣに属する音声データを生成し、各クラスタのＧＭＭを用いて学習話者をクラスタリングし直したものが格納されている。したがって、話者特徴抽出部３３は、少ない学習データから、発話内容に因らずに話者の特徴を良く表わすクラスタ情報を得ることができる。その結果、高い認識率を得ることができるのである。

【００８７】尚、本実施の形態における上記話者適応音響モデル格納部３８に格納する話者適応モデルの与え方には、上述の与え方の以外に、話者クラスタを用いる方法を採用してもよい。そして、この二通りの与え方を、音声認識装置の規模や入力音声データの量や質に応じて使い分けるのである。ここで、音声データの質とは尤度の上昇具合であり、話者特徴抽出部３３は、上記二通りの与え方による尤度の上昇具合を見計らって、上昇の大きい方法を採用するのである。長いエンロール期間が許容できる音声認識装置の場合には、このような推定処理も可能となる。尚、上記話者クラスタを用いる方法においては、教師語彙を与える上記第１実施の形態における選択法[b]に基づいて話者クラスタを選択する。そして、選択された話者クラスタの音響モデルを話者適応モデルとして話者適応音響モデル格納部３８に格納するのである。

【００８８】尚、上述した各実施の形態においては、各学習話者の音声データを声道長伸縮係数αに基づいてクラスタリングし、各クラスタのＧＭＭを用いて学習話者をクラスタリングし直したＧＭＭ格納部を搭載した音声認識装置、および、上記クラスタリングしたＧＭＭで成る話者特徴を用いて話者正規化あるいは話者適応を行う音声認識装置について説明している。しかしながら、この発明は、上記クラスタリングされたＧＭＭの何れかを話者特徴として抽出する話者特徴抽出装置にも適用されるものである。

【００８９】ところで、上記第１実施の形態における話者クラスタ選択部３，ＧＭＭ格納部６およびＧＭＭ学習部１２による上記話者特徴抽出装置としての機能は、プログラム記録媒体に記録された話者特徴抽出処理プログラムによって実現される。上記プログラム記録媒体は、ＲＯＭでなるプログラムメディアである。または、外部補助記憶装置に装着されて読み出されるプログラムメディアであってもよい。尚、何れの場合においても、上記プログラムメディアから話者特徴抽出処理プログラムを読み出すプログラム読み出し手段は、上記プログラムメディアに直接アクセスして読み出す構成を有していてもよいし、ＲＡＭ(ランダム・アクセス・メモリ)に設けられたプログラム記憶エリア(図示せず)にダウンロードして、上記プログラム記憶エリアにアクセスして読み出す構成を有していてもよい。尚、上記プログラムメディアからＲＡＭのプログラム記憶エリアにダウンロードするためのダウンロードプログラムは、予め本体装置に格納されているものとする。

【００９０】ここで、上記プログラムメディアとは、本体側と分離可能に構成され、磁気テープやカセットテープ等のテープ系、フロッピー（登録商標）ディスク，ハードディスク等の磁気ディスクやＣＤ(コンパクトディスク)－ＲＯＭ，ＭＯ(光磁気)ディスク，ＭＤ(ミニディスク)，ＤＶＤ(ディジタルビデオディスク)等の光ディスクのディスク系、ＩＣ(集積回路)カードや光カード等のカード系、マスクＲＯＭ，ＥＰＲＯＭ（紫外線消去型ＲＯＭ)，ＥＥＰＲＯＭ(電気的消去型ＲＯＭ)，フラッシュＲＯＭ等の半導体メモリ系を含めた、固定的にプログラムを担持する媒体である。

【００９１】また、上記各実施の形態における音声認識装置，音声合成装置および話者特徴抽出装置は、モデムを備えてインターネットを含む通信ネットワークと接続可能な構成を有していれば、上記プログラムメディアは、通信ネットワークからのダウンロード等によって流動的にプログラムを担持する媒体であっても差し支えない。尚、その場合における上記通信ネットワークからダウンロードするためのダウンロードプログラムは、予め本体装置に格納されているものとする。または、別の記録媒体からインストールされるものとする。

【００９２】尚、上記記録媒体に記録されるものはプログラムのみに限定されるものではなく、データも記録することが可能である。

【００９３】

【発明の効果】以上より明らかなように、第１の発明の話者特徴抽出装置は、学習話者をクラスタリングするに際して、先ず、学習話者クラスタリング手段によって、各学習話者を標準話者に対する声道長の伸縮係数αに基づいてクラスタリングし、音響モデル生成手段によって、各クラスタに属する話者集合毎にＧＭＭを生成し、再クラスタリング手段によって、上記ＧＭＭ群を用いて学習話者を再クラスタリングするので、各クラスタの初期値として声道長という大局的な特徴を明示的に与え

て、効率よくクラスタリングを行うことができる。さらに、ＧＭＭを用いて学習話者を再クラスタリングすることによって、発話内容に因らずに話者の特徴を良く表わすクラスタを得ることができ、声道長伸縮係数αの抽出誤りも修復できる。

【００９４】したがって、ループ学習手段によって、所定の条件を満たすまで上記ＧＭＭの生成と学習話者の再クラスタリングとを繰り返して得られたＧＭＭが格納された音響モデル格納部を用いて、話者クラスタ選択部によって、入力話者の音声サンプルに対して最大尤度を呈するＧＭＭを選択することによって、発話内容に因らずに精度良く入力話者の特徴を抽出することができる。

【００９５】また、１実施例の話者特徴抽出装置は、上記学習話者のクラスタリングに際して、音声サンプル豊富化手段によって、ある注目クラスタの隣接クラスタまたは近傍クラスタに属する話者の音声サンプルに対して周波数伸縮を行い、上記注目クラスタに属する音声サンプルを生成して編入するので、学習データ不足を補って、少ない発声データからでも各クラスタの音響モデルを精密に構築できる。したがって、少ない発声データでより精度良く入力話者の特徴を抽出できる。

【００９６】また、１実施例の話者特徴抽出装置は、上記音声サンプル豊富化手段によって上記学習話者の音声サンプルに対して周波数伸縮を行う音声区間を、有音・無音の別および調音点に基づいて限定するので、声道長の差の影響を受け難い音素や無音部を上記周波数軸伸縮の対象外にして、声道長の差の影響を受け難い音素や無音部まで変形されることを防止できる。

【００９７】また、１実施例の話者特徴抽出装置は、上記再クラスタリング手段によって上記学習話者を再クラスタリングする場合に、上記再クラスタリングの対象となる対象学習話者が再クラスタリングの前後に属しているクラスタの伸縮係数αが所定値以上離れている場合には、当該対象学習話者を上記再クラスタリングの対象から外すので、声道長伸縮係数αが極端に異なる話者同士が同じクラスタに属することを防止できる。

【００９８】また、１実施例の話者特徴抽出装置は、サブクラスタ生成手段によって、上記ループ学習手段によって最終的にクラスタリングされた各クラスタに属する学習話者を更にクラスタリングしてサブクラスタを生成し、上記各サブクラスタに属する話者集合毎にＧＭＭを生成し、上記音響モデル格納部は、上記サブクラスタ生成手段によって生成されたＧＭＭの群を、各サブクラスタの伸縮係数αに対応付けて格納するので、このサブクラスタによって声道長以外の要因にも対応することができ、より精密な話者特徴を抽出できる。

【００９９】また、第２の発明の話者特徴抽出装置は、標準話者に対する声道長の伸縮係数αに基づいて学習話者をクラスタリングし、各クラスタ毎のＧＭＭの生成とそのＧＭＭ群を用いた学習話者の再クラスタリングとを所定の条件を満たすまで繰り返し、最終的に生成されたＧＭＭ群を話者クラスタ毎に格納した音響モデル格納部を用いて、話者クラスタ選択部によって、入力話者の音声サンプルに対して最大尤度を呈するＧＭＭを選択して入力話者の特徴とするので、発話内容に拘らず精度良く入力話者の特徴を抽出することができる。

【０１００】また、第３の発明の音声認識装置は、上記第１の発明あるいは第２の発明の話者特徴抽出装置の音響モデル格納部における話者クラスタ毎に生成されたＨＭＭの群が格納されたＨＭＭ格納部から、切換部によって、上記話者特徴抽出装置で選択された話者クラスタのＨＭＭを音声認識用の音響モデルとして切り換え選出するので、発話内容に拘らずに少ない音声データで、上記入力話者の特徴を精度良く表わすＨＭＭを用いて、入力話者の音声を正確に認識することができる。

【０１０１】また、第４の発明の音声認識装置は、上記第１の発明あるいは第２の発明に係る音響モデル格納部における話者クラスタ毎に生成されたＨＭＭの群が格納されたＨＭＭ格納部を用いて、切換部によって、最大の尤度を呈するＨＭＭを音声認識用の音響モデルとして切り換え選出するので、発話内容に拘らずに少ない音声データで、上記入力話者の特徴を精度良く表わすＨＭＭを用いて、入力話者の音声を正確に認識することができる。

【０１０２】また、第５の発明の音声認識装置は、話者正規化手段を、上記第１の発明あるいは第２の発明の話者特徴抽出装置と、入力話者の音声サンプルに基づいて上記話者特徴抽出装置によって選択されたＧＭＭに対応する声道長の伸縮係数αを用いて、上記入力音声のスペクトルの周波数軸を伸縮する周波数ワープ手段で構成したので、発話内容に拘らずに少ない音声データで、上記入力話者の特徴を精度良く表わす声道長の伸縮係数αを用いて、より標準話者の音響特徴量に近づくように話者正規化を行うことができる。したがって、高い音声認識率を得ることができる。

【０１０３】また、第６の発明の音声認識装置は、話者適応手段を、上記第１の発明あるいは第２の発明の話者特徴抽出装置と、入力話者の音声サンプルに基づいて上記話者特徴抽出装置によって選択されたＧＭＭに対応する声道長の伸縮係数αの逆数を用いて、音響モデルのスペクトルの周波数軸を伸縮する周波数ワープ手段で構成したので、発話内容に拘らずに少ない音声データで、上記入力話者の特徴を精度良く表わす声道長の伸縮係数αの逆数を用いて、より入力話者の音響特徴量に近づくように話者適応を行うことができる。したがって、高い音声認識率を得ることができる。

【０１０４】また、第７の発明の話者特徴抽出方法は、各学習話者を標準話者に対する声道長の伸縮係数αに基づいてクラスタリングし、各クラスタに属する話者集合毎にＧＭＭを生成し、上記ＧＭＭ群を用いて学習話者を

27

再クラスタリングするので、各クラスタの初期値として声道長という大局的な特徴を明示的に与えて、効率よくクラスタリングを行うことができる。さらに、ＧＭＭを用いて学習話者を再クラスタリングすることによって、発話内容に因らずに話者の特徴を良く表わすクラスタを得ることができ、声道長伸縮係数αの抽出誤りも修復できる。

【０１０５】したがって、所定の条件を満たすまで上記ＧＭＭの生成と学習話者の再クラスタリングとを繰り返して得られたＧＭＭを格納した音響モデル格納部を用いて、入力話者の音声サンプルに対して最大尤度を呈するＧＭＭを選択することによって、発話内容に因らずに精度良く入力話者の特徴を抽出することができる。

【０１０６】また、第８の発明のプログラム記録媒体は、コンピュータを、上記第１の発明に係る学習話者クラスタリング手段,音響モデル生成手段,再クラスタリング手段,ループ学習手段,音響モデル格納部および話者クラスタ選択部として機能させる話者特徴抽出処理プログラムを記録しているので、上記第１の発明の場合と同様に、発話内容に因らずに精度良く入力話者の特徴を抽出することができる。

【図面の簡単な説明】

【図１】　この発明の話者特徴抽出装置を搭載したクラスタリング方式による音声認識装置におけるブロック図である。

28

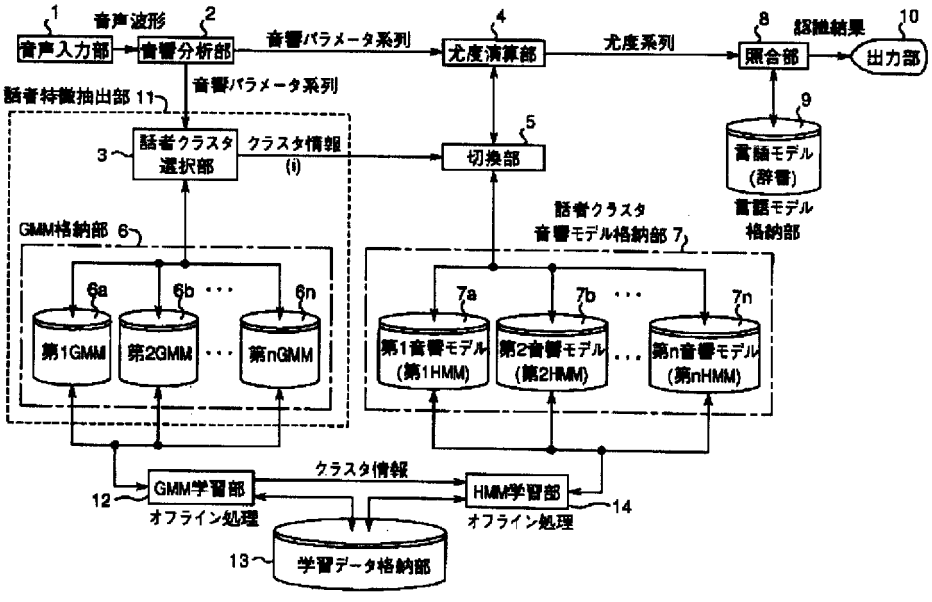
【図２】　図１とは異なる話者正規化方式による音声認識装置におけるブロック図である。

【図３】　図１および図２とは異なる話者適応方式による音声認識装置におけるブロック図である。

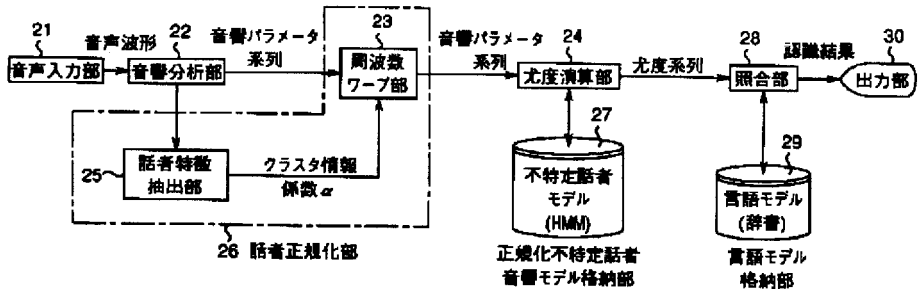
【符号の説明】

１，２１，３１…音声入力部、
２，２２，３２…音響分析部、
３…話者クラスタ選択部、
４，２４，３７…尤度(音韻類似度)演算部、
５…切換部、
６…ＧＭＭ格納部、
７…話者クラスタ音響モデル格納部、
８，２８，３９…照合部、
９，２９，４０…言語モデル格納部、
１０，３０，４１…出力部、
１１，２５，３３…話者特徴抽出部、
１２…ＧＭＭ学習部、
１３…学習データ格納部、
１４…ＨＭＭ学習部、
２３，３４…周波数ワープ部、
２６…話者正規化部、
２７，３５…正規化不特定話者音響モデル格納部、
３６…話者適応部、
３８…話者適応音響モデル格納部。

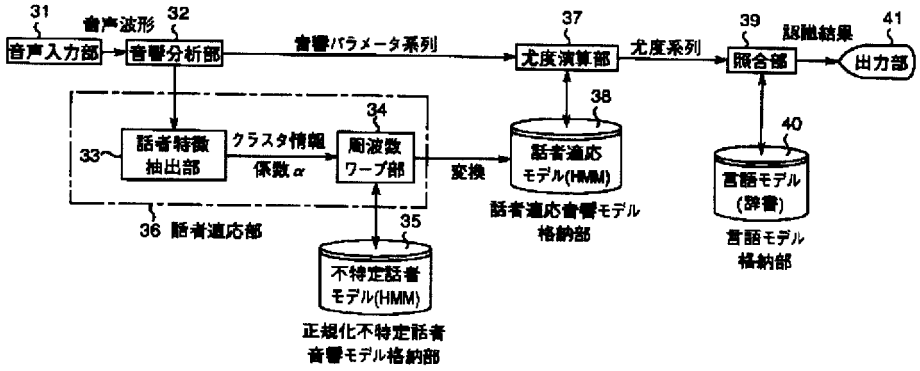
【図１】



【図２】



【図３】



フロントページの続き

(51) Int. Cl.[7]	識別記号	F I	テーマコード（参考）
G10L 15/14		G10L 3/00	537A
15/18			531K
			535C